Unsupervised Human Activity Recognition on the UCI-HAR Dataset

Eduard-Valentin Dumitrescul

January 6, 2026

Contents

1 Introduction

Human Activity Recognition (HAR) plays a significant role in modern health monitoring, fitness tracking, and elderly care systems. This technology utilizes inertial sensors—specifically accelerometers and gyroscopes—embedded in mobile devices to gather high-frequency data about a user's movement. By training machine learning models on this data, it is possible to classify physical movement into distinct categories.

Traditional learning models used for HAR, such as **Support Vector Machines (SVM)**, are supervised. These models require large, manually labeled datasets to function effectively. However, labeling data is a labor-intensive process that is not always feasible for real-world applications.

This project explores **unsupervised learning**, specifically **Self-Organizing Maps (SOM)** and **Mean-Shift Clustering**. The primary goal is to determine if these algorithms can discover the underlying structure of human movement without prior knowledge of activity labels. A supervised SVM is implemented alongside these models to serve as a performance benchmark.

1.1 UCI-HAR Dataset Overview

The **UCI-HAR** dataset [1] is a standard benchmark for Human Activity Recognition tasks. The data was collected from 30 subjects performing 6 activities (walking, walking upstairs, walking downstairs, sitting, standing, laying)

During the experiments, the subjects wore a Samsung Galaxy S II smartphone tied to their waist. Using the device's embedded accelerometer and gyroscope, movement and orientation data was collected at a rate of 50Hz The captured information was later manually labeled using video recording of the subjects.

The resulting database consists of 7352 train examples and 2947 test examples were created. Each example consists of a 561-feature vector derived from the collected data.

2 Methodology

The UCI-HAR dataset presents an opportunity to employ two distinct learning paradigms: supervised and unsupervised. The primary objective is to compare these two types of algorithms in order to discover if the models that do not have access to class labels are able to discoer the underlying structure of the data.

2.1 Models

A total of 3 models (1 supervised, 2 unsupeerised) have been trained and evaluated on the UCI-HAR dataset.

2.1.1 Support Vector Machine

For the benchmark model, a **Support Vector Machine** is used. It is a classic machine learning algorithm known for its efficiency and high-performance on tabular data. In the

context of UCI-HAR, the SVM is effective in finding hyperplanes delimiting the activity classes.

2.1.2 Self-Organizing Maps (SOM)

The first unsuperised model is a **Self-Organizing Map (SOM)**. This is a type of artificial neural network that performs dimensionality-reduction by mapping high-dimensional data onto a low-dimensional space (often 2D), while preserving the topological properties of the input data. Because similar items are clustered together on the grid, the overall structure can be easily visualized.

2.1.3 Mean-Shift Clustering

The second unsupervised model is a **Mean-Shift Clustering** algorithm, which clusters data by iteratively shifting items towards high-density areas.Unlike other algorithms, Mean-Shift Clusterin determines the number of clusters based on the data itself, allowing us to see if the data naturally groups together.

2.2 Model Input Strategies

While selecting the feature sets, it is important to consider that clustering algorithms, specifically Self-Organizing Maps (SOM) and Mean-Shift Clustering, suffer from the **curse of dimensionality**. A high number of dimensions causes the data point to be sparse, which makes distance-based metrics (like Euclidean distance) unreliable. To mitigate this, the models will be trained on multiple feature sets:

- The original 561-feature vector. This serves as a baseline for the models
- Manually selected features. A reduced set of features based on physical reasoning.
- Principal Component Analysis (PCA): A reduced set of uncorrelated components, which preserves most of the information.

2.3 Supervised Model Evaluation

2.4 Unsupervised Model Evaluation

2.4.1 Internal Validation (Clustering Specific Metrics)

2.4.2 External Validation (Projected Label Evaluation)

3 Feature Extraction and Engineering

3.1 Feature Set 1: Original 561-feature vectors

The baseline feature set consists of the full 561-dimensional vectors provided by the UCI-HAR dataset. They were calculated from the 3-axial linear acceleration and angular velocity signals. The creators applied various filters and transformations, including Fast Fourier Transform to produce frequency domain variabled[1]. This set represents the most comprehensive data available, but poses the risk of the curse of dimensionality.

3.2 Feature Set 2: Manually Selected Domain Features

A hand-picked set of 15 features was selected based on the physical state differences between the 6 activities.

- **Static Orientation (Laying vs. Others):** To distinguish the "Laying" state (horizontal) from vertical states (Sitting, Standing, Walking), we selected the mean gravity acceleration components across all three axes: `tGravityAcc-mean()-X`, `Y`, and `Z`.
- **Dynamic vs. Static States (Moving vs. Non-Moving):** To separate active movements (Walking, Stairs) from rest (Sitting, Standing, Laying), we included the Signal Magnitude Area (`tBodyAcc-sma()`) and the standard deviation of body acceleration magnitude (`tBodyAccMag-std()`).
- **Rhythmic Motion and Gait (Walking vs. Stairs):** Distinguishing walking from stair navigation requires capturing the cadence and intensity of movement. We have chosen the following features in an attempt to capture that information.
  - Gyroscope Jerk standard deviation: `tBodyGyroJerk-std()-X`, `Y`, and `Z`.
  - Mean angular velocity: `tBodyGyro-mean()-X`.
  - Frequency-domain body acceleration: `fBodyAcc-mean()-X`.
  - Frequency-domain magnitude mean frequency: `fBodyAccMag-meanFreq()`.
- **Postural Stability (Standing vs. Sitting):** These two states are predicted to be difficult to separate. To capture the subtle movements present while standing compared to the more rigid seated posture, we included:
  - Gyroscope standard deviation: `tBodyGyro-std()-X`, `Y`, and `Z`.
  - Angular velocity jerk magnitude: `tBodyGyroJerkMag-std()`.

3.3 Feature Set 3: Principal Component Analysis (PCA)

Principal Component Analysis is used to transform the original 561-dimensional vectors into lower-dimensional ones, while preserving the majority of the information. This algorithm calculates new uncorrelated features that will be fed to the models.

Various numbers of components have been extracted using PCA:

- **10 components** preserve 80% of the original dataset's variance.

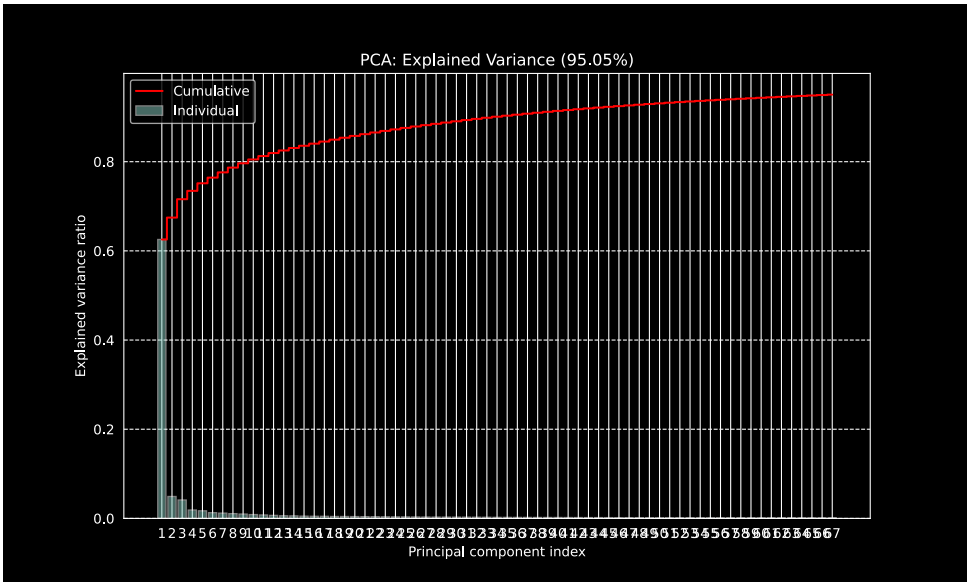

Figure 1: PCA Explained Variance Ratio

- **34 compoentns** preserve 90% of the original dataset's variance.
- **67 components** preserve 95% of the original dataset's variance.

3.4 Training on Full Feature Set

Grid search was used to find the best parameter configuration for the SVM. Results can be seen in the table below. The model achieves **96.54%** accuracy on the test subset.

	Regularization Parameter (C)			
Gamma (γ)	**0.1**	**1**	**10**	**100**
`'scale'`	91.28%	95.05%	96.20%	**96.54%**
`0.01`	92.30%	94.98%	96.20%	**96.54%**
`0.001`	80.96%	92.77%	95.39%	96.17%

Table 1: Grid Search Results for SVM on Full Feature Set (561 features). Results represent Test Accuracy using an RBF kernel.

Activity	**Precision**	**Recall**	**F1-Score**	**Support**
Walking	0.96	0.99	0.97	496
Walking Upstairs	0.95	0.96	0.96	471
Walking Downstairs	0.99	0.95	0.97	420
Sitting	0.98	0.90	0.94	491
Standing	0.92	0.98	0.95	532
Laying	1.00	1.00	1.00	537
Accuracy			**0.97**	**2947**
Macro Avg	0.97	0.96	0.96	2947
Weighted Avg	0.97	0.97	0.97	2947

Table 2: Detailed Classification Report for the Optimized SVM (Full Feature Set)

3.5 Training on Manually Selected Feature Set

To evaluate the efficiency of the domain-specific features, the SVM was retrained on the 15-dimensional subset. As shown in the results below, the model remains highly effective despite a significant reduction in input dimensionality.

	Regularization Parameter (C)			
Gamma (γ)	**0.1**	**1**	**10**	**100**
`'scale'`	89.48%	**90.16%**	88.97%	89.01%
`0.01`	85.31%	89.89%	89.92%	89.38%
`0.001`	51.95%	84.97%	89.55%	89.82%

Table 3: Grid Search Results for SVM on Manual Feature Set (15 domain-specific features). Results represent Test Accuracy using an RBF kernel.

Activity	Precision	Recall	F1-Score	Support
Walking	0.91	0.96	0.93	496
Walking Upstairs	0.87	0.89	0.88	471
Walking Downstairs	0.94	0.85	0.89	420
Sitting	0.89	0.77	0.83	491
Standing	0.82	0.91	0.86	532
Laying	1.00	1.00	1.00	537
Accuracy			**0.90**	**2947**

Table 4: Classification Report: Manual Feature Set.

3.6 Training on PCA Feature Set (80% Variance)

To evaluate the impact of automated dimensionality reduction, the SVM was trained on the first 10 principal components, which capture 80% of the dataset's variance. This configuration tests the model's robustness when restricted to the most significant statistical features.

The grid search results in Table 5 indicate that the model achieves a peak accuracy of **88.26%** at $C = 100$ and $\gamma = scale$. While lower than the manual feature set (90.16%), it represents a high degree of information density for only 10 dimensions.

	Regularization Parameter (C)			
Gamma (γ)	**0.1**	**1**	**10**	**100**
`'scale'`	86.60%	87.72%	88.12%	**88.26%**
`0.01`	85.71%	87.72%	87.75%	88.12%
`0.001`	78.08%	86.02%	87.72%	88.02%

Table 5: Grid Search Results: PCA Feature Set (10 components, 80% Variance).

Activity	Precision	Recall	F1-Score	Support
Walking	0.87	0.98	0.92	496
Walking Upstairs	0.90	0.90	0.90	471
Walking Downstairs	0.93	0.80	0.86	420
Sitting	0.81	0.72	0.76	491
Standing	0.77	0.87	0.82	532
Laying	1.00	0.98	0.99	537
Accuracy			**0.88**	**2947**

Table 6: Classification Report: PCA Feature Set (80% Variance).

3.7 Interpretation

The SVM achieves great accuracy on the full feature dataset and good accuracy on the manually selected features and PCA feature set. The main difficulty encounted by the model is the distunguishing between *Sitting* and *Standing*, as expected (3, 4 6). *Laying* is the easiest to classify, achieving almost perfect metrics for all three features sets. Static and dynamic classes are easily distinguished between one another.

4 Unsupervised Model 1: Self-Organizing Maps

Self-Organizing Maps (SOM) were employed to projects the high-dimensional data (561, 15, repectively 10 dimensions) to a 2-dimensional grid.

Model Parameters:

- **Grid size**: Determines the resolution of the final map and the number of clusters that represent the input space.
- **Learning rate**: Controls how aggresive the weihts are updated between iterations.
- **Sigma**: Defines the neighborhood radius. A larger value prioritizes global ordering, while a lower one focuses on local fine-tuning

The SOM algorithm works by initializing a weight vector for each neuron in the grid. Each iteration consists of three steps:

1. **Competition**: For each input sample, the neuron with the smallest Euclidean distance is calculated (*Best Matching Unit*).
2. **Cooperation**: The BMU identifies the neurons in its neighborhood.
3. **Adaptation**: The weight vectors of the BMU and its neighbors are updated.

The two most important metrics of the SOM are:

- **Quantization Error (QE)**: The average Euclidean distance between each input sample and its BMU.
- **Topographic Error (TE)**: The proportion of input samples for each the 1st BMU and the 2nd BMU are not neighbors

4.1 Training

Grid Search was emplyed on each feature set with the following parameters:

- **Grid Size(4x4, 6x6, 8x8, 10x10, 20x20)**
- **Sigma Ratio(0.2, 0.4, 0.6)**: This is relative to the grid size (e.g. sigma ratio 0.2 for grid size 10x10 is sigma 2)
- **Learning Rate(0.5)**: For fast learning
- **Number of Iterations(5000)**

Quantization Error (QE) and **Topographic Error (TE)** are calculated for each configuration. For ensuring the input data is correctly mapped to the grid, only the results having a TE lower than 0.1 are considered. Then, the QE is the deciding factor, while also prioritizing small grid sizes.

Grid Size	Sigma 0.2	Sigma 0.4	Sigma 0.6
4 × 4	4.277 / 0.511	4.375 / 0.196	4.557 / 0.084
6 × 6	3.879 / 0.384	4.276 / 0.119	4.461 / 0.048
8 × 8	3.789 / 0.266	4.206 / 0.093	4.383 / 0.067
10 × 10	3.690 / 0.245	4.158 / 0.090	4.400 / 0.159
20 × 20	3.559 / 0.143	4.141 / 0.069	4.309 / 0.055

Note: Cells show Quantization Error (QE) / Topographic Error (TE)

Table 7: SOM Grid Search Results: Full (561) Feature Set

Grid Size	Sigma 0.2	Sigma 0.4	Sigma 0.6
4 × 4	1.510 / 0.190	1.613 / 0.034	1.775 / 0.007
6 × 6	1.274 / 0.279	1.488 / 0.030	1.694 / 0.035
8 × 8	1.155 / 0.231	1.455 / 0.040	1.700 / 0.028
10 × 10	1.067 / 0.101	1.422 / 0.048	1.614 / 0.049
20 × 20	0.977 / 0.072	1.381 / 0.068	1.579 / 0.037

Note: Cells show Quantization Error (QE) / Topographic Error (TE)

Table 8: SOM Grid Search Results: Manual (15) Feature Set

Grid Size	Sigma 0.2	Sigma 0.4	Sigma 0.6
4 × 4	2.253 / 0.167	2.453 / 0.078	2.755 / 0.150
6 × 6	1.897 / 0.205	2.357 / 0.062	2.578 / 0.065
8 × 8	1.794 / 0.108	2.261 / 0.078	2.488 / 0.051
10 × 10	1.726 / 0.190	2.211 / 0.055	2.463 / 0.059
20 × 20	1.621 / 0.088	2.144 / 0.036	2.401 / 0.037

Note: Cells show Quantization Error (QE) / Topographic Error (TE)

Table 9: SOM Grid Search Results: PCA (80%) Feature Set

As observed from the three tables above, a good configuration for the SOM requires a balance between topographic preservation and quantization accuracy. Specifically, configurations with a *Sigma Ratio* of 0.2 often resulted in a Topographic Error exceeding the 0.1 threshold, indicating a neighborhood radius too small to maintain the global structure of the map.

By prioriziting compact grids and the selection criteria (filtering for $TE < 0.1$ and selecting the minimum QE):

- **Full (561):** The 6×6 grid with **0.6 Sigma Ratio** ($QE = 4.461, TE = 0.048$).
- **Manual (15):** The 4×4 grid with **0.4 Sigma Ratio** ($QE = 1.488, TE = 0.035$).
- **PCA (80%):** The 6×6 grid with **0.4 Sigma Ratio** ($QE = 2.357, TE = 0.062$).

The table below shows the results after running the previous 3 configuration on both 4 by 4 and 6 by 6 grids. Three new metrics were added:

- **Accuracy**: Each cell was labeled it using a majority vote algorithm based on the labels for the samples belonging to it. Then, using the testing subset, the accuracy was calculated.
- **Average Purity**: Each cell has an associated purity based on what percentage of the samples beloging to it are labeled like the majority.
- **Pure Neurons**: The percentage of cells having the purity greater than 90%.

Feature Set	Grid	ACC	QE / TE	Avg. Purity	Pure Neurons
PCA (80%)	6×6	**0.7475**	2.448 / 0.051	**0.8042**	30.56%
	4×4	0.6098	2.805 / 0.150	0.7714	31.25%
Full (561)	6×6	0.6583	4.675 / 0.154	0.8038	**47.22%**
	4×4	0.6298	4.906 / 0.279	0.7762	43.75%
Manual (15)	6×6	0.6922	**1.661 / 0.059**	0.7484	27.78%
	4×4	0.6230	1.834 / **0.041**	0.7187	25.00%

Table 10: Comparison of SOM Performance across Feature Sets and Grid Sizes

From these results it can be concluded that, while being far from the results of a supervised SVM, the SOM is able to successfully cluster and distinguish activity labels, without having access to labels during training. The best accuracy and average purity is achieved when training on the PCA feature set with a 6 by 6 grid. However, the manual set obtains the smallest Quantization Error and Topographic Error, while having a larger input dimension (15 compared to 10 of the PCA feature set).

4.2 Notes

It is important to acknowledge that a grid with more cells will naturally achieve better results, due to how they are labeled. For this reason, The 4 by 4 manual feature set model might perform better than the 6 by 6 PCA model. Due to the discrepancy between the unsupervised nature of the models and the supervised nature of the task, it is hard to selected an objectively best model.

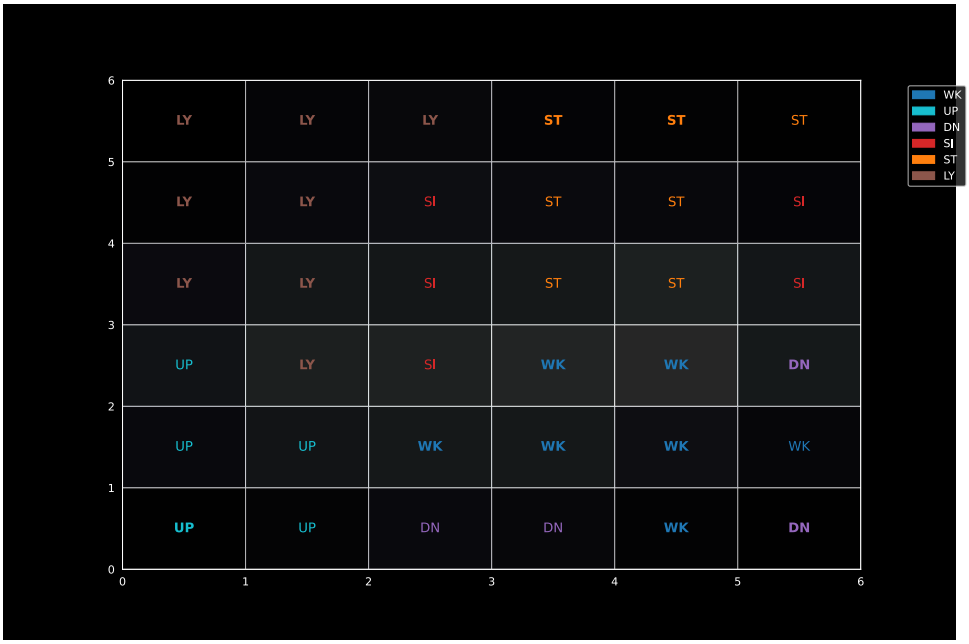

Figure 2: Labeled grid for the 6 by 6 PCA model

4.3 Results Evaluation (Visualizing the 40x40 Activity Map)

5 Unsupervised Model 2: Mean-Shift Clustering

5.1 Training on Manually Selected Feature Set

5.1.1 Bandwidth Estimation and Density Tuning

5.2 Results Evaluation (Cluster Purity and Recall)

5.3 Training on PCA Feature Set

5.3.1 Bandwidth Estimation and Density Tuning

5.4 Results Evaluation (Silhouette Analysis)

6 Comparative Analysis and Discussion

7 Conclusion

References

[1] Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. *Human Activity Recognition Using Smartphones*. UCI Machine Learning Repository. 2013. DOI: 10.24432/C54S4K.

A Supplementary Figures

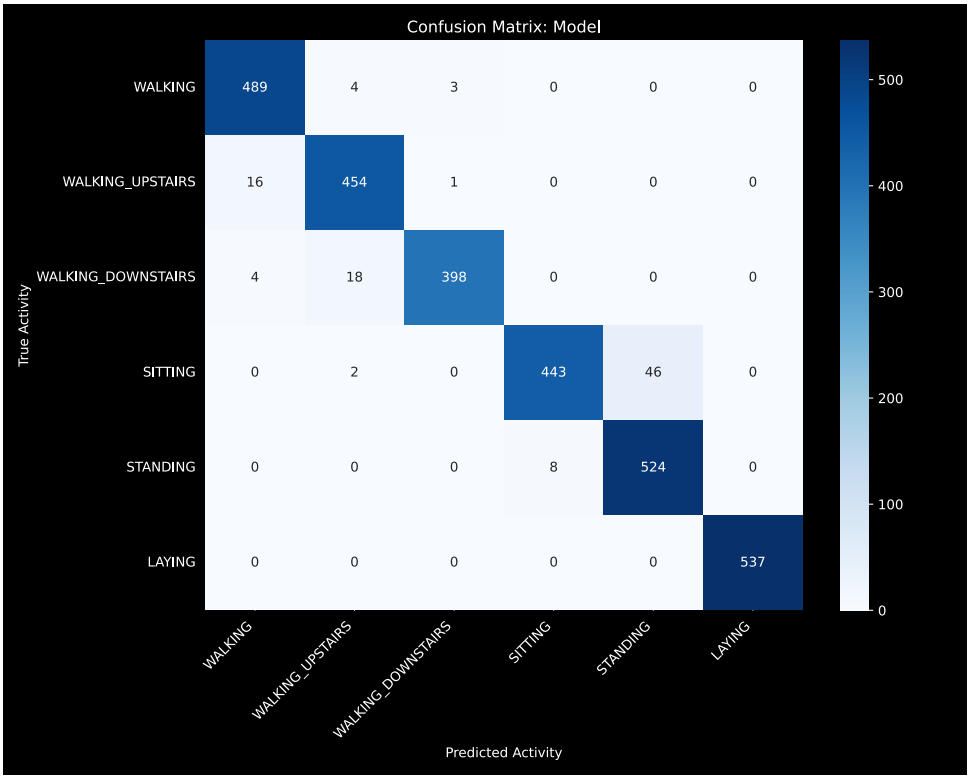

Figure 3: Full Confusion Matrix: 561-Feature SVM.

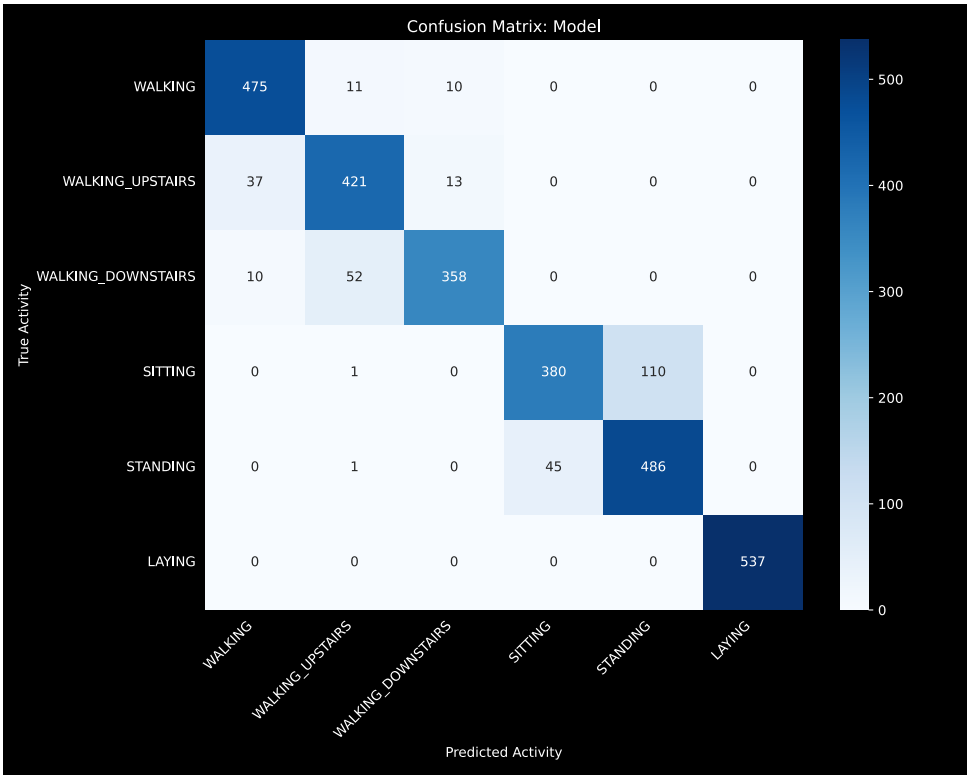

Figure 4: Full Confusion Matrix: 15-Feature Manual SVM.

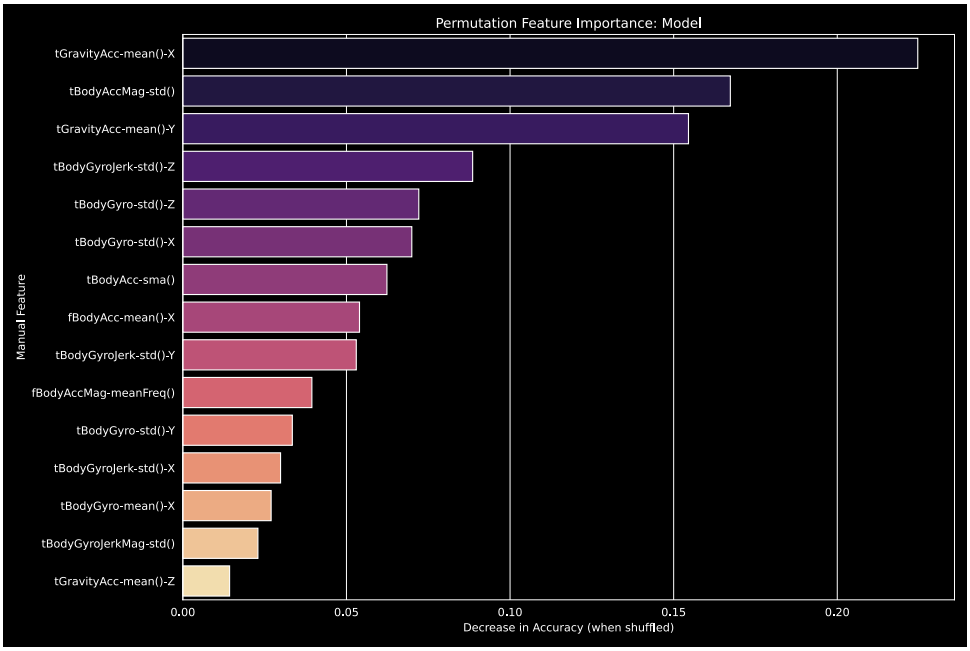

Figure 5: Permutation Importance for Manual Feature Set.

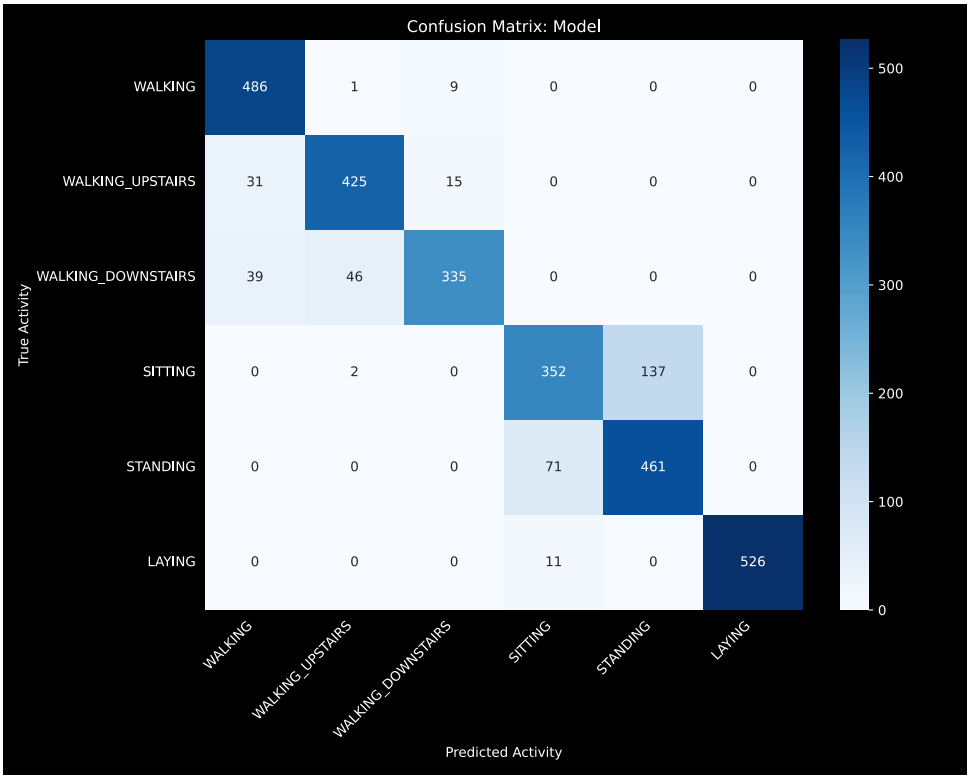

Figure 6: Full Confusion Matrix: 10-Feature PCA SVM.